IEEE International Conference on Multimedia & Expo 2001, Tokyo, Japan, Aug. 2001.

MOTION COMPENSATED LIFTING WAVELET AND ITS APPLICATION IN VIDEO CODING

Lin Luo[1], Jin Li[3], Shipeng Li[2], Zhenquan Zhuang[1] and Ya-Qin Zhang[2]

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China.
[2]Microsoft Research China, Sigma Ctr. 49# Zhichun Rd, Haidian, Beijing, P.R.China
[3] Microsoft Research, Signal Processing, One Microsoft Way, Bld. 113/3033, Redmond WA 98052
e-mail: luolynn@hotmail.com, {jinl, spli}@microsoft.com

ABSTRACT

A motion compensated lifting (MCLIFT) framework is proposed for the 3D wavelet video coder. By using bi-directional motion compensation in each lifting step of the temporal direction, the video frames are effectively de-correlated. With proper entropy coding and bitstream packaging schemes, the MCLIFT wavelet video coder can be scalable in frame rate and quality level. Experimental results show that the MCLIFT video coder outperforms the 3D wavelet video coder with the same entropy coding scheme by an average of 1.1-1.6dB, and outperforms MPEG-4 coder by an average of 0.9-1.4dB.

1. INTRODUCTION

Wavelet transform is an effective tool for image/video decomposition. It packs the energy of the image/video into a small set of wavelet coefficients, which are further compressed by entropy coding. An additional advantage of the wavelet transform is that a lower resolution signal is generated with each level of wavelet decomposition. This leads to a nice scalability, i.e., a subset of the coefficients can be accessed and decoded to obtain a lower resolution image/video. Recently, a wavelet based image coding standard - JPEG 2000 [17] has been proposed. Compared with JPEG, the current DCT-based image compression standard, JPEG 2000 offers not only offers superior compression performance, but also offers scalabilities in both resolution and quality level which are very useful in the consumer and the Internet applications.

Applications of the wavelet in video coding follow two paths: motion compensated wavelet residue coding [9]-[13] and the 3D wavelet video coding [1]-[8]. In the motion compensated wavelet residue coder, the current frame is predicted by the content from the previous frame, subject to the object motion. The prediction residue is then further encoded by a wavelet coder. The framework of the coder is very similar to the existing video coding standard, such as the MPEG, except that the residue coder is a wavelet coder instead of a DCT coder. On the other hand, the 3D wavelet coder applies the wavelet transform in all three directions, i.e., the temporal, horizontal and vertical directions, and then encodes the transformed coefficients as a whole using entropy coder. With a proper entropy coding and bitstream packaging scheme, a 3D wavelet compressed bitstream can achieve quality and temporal scalabilities simultaneously, which is a very desirable feature in video delivery and storage. However, with moderate-motion sequences, the current 3D wavelet coder underperforms the existing state-of-the-art video coding standard, such as the MPEG-4[7][8].

The primary weakness of the existing 3D wavelet video coder lies in the temporal filter. Object motion (such as panning and zooming) in the video causes the object to be misaligned along the temporal direction, and leads to inefficiency in wavelet decomposition in the temporal direction. Works have been done to improve the correlation of video signal along the temporal direction. Taubman and Zakhor[4] pan shifted the video sequence before the 3D wavelet transform was applied. Wang et al. [7] proposed to register and warp all image frames into a common coordinate system and then apply a 3D wavelet transform with an arbitrary region of support to the warped volume. To make use of the local block motion, Ohm [5] incorporated block matching and carefully handled the covered/uncovered, connected/unconnected regions. By trading off the invertibility requirement, Tham et al. [6] employed a block-based motion registration for the low motion sequences without filling the holes caused by individual block motion. A threading approach has been proposed by Xu et. al. [8] so that the pixel along the same motion trajectory is aligned for wavelet filtering. In the above mentioned approaches, either the much simpler global panning/warping motion model was used, or when a block motion model was used, the temporal decomposition schemes become very complex. Besides, even after the motion alignment, the compression efficiency of the 3D wavelet coder is still not satisfactory.

In this paper, a motion compensated lifting (MCLIFT) scheme is proposed for temporal wavelet filtering. MCLIFT applies multi-level lifting operation in the temporal direction, with each elementary lifting operation being a bi-directional motion compensated prediction. After temporal MCLIFT operation, the decomposed frames are transformed by a 9-7 bi-orthogonal wavelet filter within the frame. The transformed coefficients are then entropy encoded. Experimental results show that the MCLIFT video coder outperforms MPEG-4 by 0.9-1.4dB on average.

The paper is organized as follows. The framework of the MCLIFT video coder, including the temporal MCLIFT filter and the coding of the wavelet coefficients, is proposed in Section 2. Experimental results are presented in Section 3. Conclusions are drawn in Section 4.

2. MOTION COMPENSATED LIFTING WAVELET

The framework of the motion compensated lifting (MCLIFT) wavelet coder can be shown in Fig. 1. A sequence of video frames are fed into the MCLIFT coder, which first decomposes the video temporarily through the MCLIFT filter, and then decomposes the video horizontally and vertically within frames by a multi-level bi-orthogonal 9-7 filter. The decomposed coeffi-

cients are then entropy encoded through a highly efficient bit-plane coder.

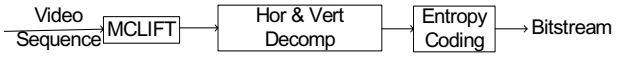

Figure 1 Framework of the motion compensated lifting wavelet (MCLIFT) video coder.

We'll examine each MCLIFT building block in details below.

2.1 Motion compensated lifting wavelet

Lifting is a memory and computationally efficient implementation of the wavelet transforms [14]. Every FIR wavelet filter can be factored into a few lifting steps [15]. A sample one-level forward and backward one-dimensional bi-orthogonal 5-3 lifting wavelet is illustrated in Figure 2. The original data $x_0, x_1, \cdots, x_6$ are input at the left, while the decomposed wavelet coefficients are output at the right two columns. It is observed that the wavelet coefficients are calculated through two stages of computation, each of which involves only half of the nodes. The lifting process can be formulated as follows:

$$\begin{cases} H_i = x_{2i+1} + a \times (x_{2i} + x_{2i+2}) \\ L_i = x_{2i} + b \times (x_{i-1} + x_i) \end{cases}, \text{ where } a = -1/2,\ b = 1/4.$$

Because each elementary forward lifting unit can be straightforwardly inversed to an inverse lifting unit, the inverse wavelet lifting can be easily derived by directly inverting the data flow of the forward one, as shown in the right side of Figure 2. Multi-level wavelet transform can be obtained by further performing the lifting wavelet filter on decomposed subbands.

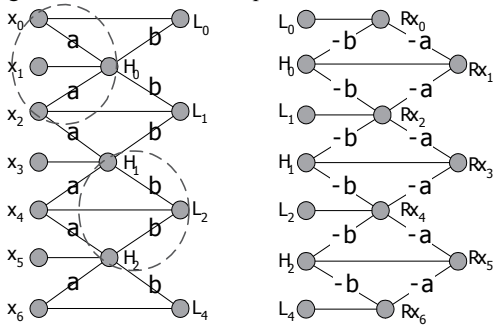

Figure 2 Forward and inverse wavelet transforms via lifting and the elementary lifting units (circled in the forward structure).

In the MCLIFT video coder, the input data are video frames. Moreover, the motion compensation is applied on the prediction branch during the temporal frame lifting operation. An elementary motion compensated temporal lifting operation is very similar to a B frame prediction operation in MPEG, and can be depicted in Figure 3. Let *B* be the current frame to be lifted, and *A* and *C* be the two neighbor (reference) frames. The motion compensated lifting (MCLIFT) operation can be formulated as:

$$MCLIFT(B) = B - a\,[\,MCP(A) + MCP(C)\,],$$

where *a* is a lifting parameter, and MCP(*X*) is the motion compensated prediction of the reference *X* frame. In the current MCLIFT implementation, we adopt block motion compensation model with half-pixel accuracy. However, advanced motion compensation schemes developed in MPEG-4 and H.26L, such as the overlapped block motion compensation, motion compensation with quarter pixel accuracy, can be used in MCLIFT as well.

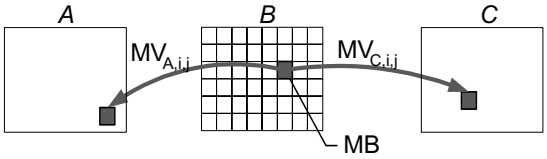

Figure 3 The elementary motion compensated lifting operation.

The 3-level temporal frame structure of the MCLIFT video coder is shown in Figure 4. Let F0, F1,⋯,F8 be 9 original video frames. We note that the frame F0 and F8 are not lifted. They are called category A frames. Frame F4 is lifted with reference to frame F0 and F8 (two category A frames), it is called a category B frame. Likewise, frame F2 and F6 is lifted with reference to two nearby category B and A frames, we call them the category C frames. Frame F1, F3, F5 and F7 are category D frames, which are lifted with reference to nearby category A, B, C frame. The transform structure can be considered as a 3-level wavelet decomposition with truncated 5-3 lifting operation, where the category D, C, B frames are calculated at the 1st, 2nd and 3rd levels, respectively. The truncated 5-3 lifting applies only the high pass lifting operation of Figure 2 with parameter $a=-1/2$, while the 2nd stage (low pass) lifting operation is not applied. The original frames, rather than the decoded frames, are used as reference in the MCLIFT operation. We have tested the full 5-3 motion compensated lifting, where the 2nd stage (low pass) lifting operation is applied. More complicated bi-orthogonal 9-7 filter has been tested for MCLIFT operation as well. However, experimental results show that both configurations perform inferior to the current one. Therefore, the truncated 5-3 MCLIFT with only the high pass lifting operation is used through the rest of the paper.

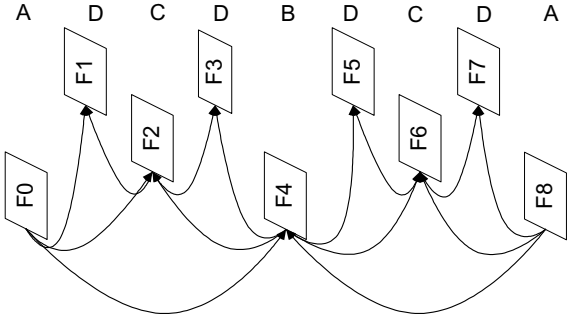

Figure 4 3-level MCLIFT temporal wavelet decomposition.

At a single level, the MCLIFT is very similar to the B frame prediction used in MPEG. However, the MCLIFT differs from MPEG when multi-level MCLIFT is used. We show a comparative MPEG frame structure in Figure 5. The GOP structure is "IBBBPBBB", which is the most similar MPEG structure compared to the 3-level MCLIFT coder. Again, there are altogether 9 frames: F0, F1, ⋯, F8. Frames F0 and F8 are independently encoded as I frames. Frame F4 is encoded as a P frame with reference to frame F0. Frames F1, F2, F3, F5, F6, F7 are encoded as B frames with reference to nearby I and P frames.

Comparing with the temporal frame structure of MCLIFT (Figure 4) versus that of MPEG (Figure 5), we notice that MCLIFT provides B frame prediction for frame F4, and provides the frames F1, F3, F5 and F7 with closer reference frames F2 and F6. This improves the effectiveness of temporal decomposition of MCLIFT. Both MCLIFT and MPEG can easily scale the

decoded video down to 1/2, 1/4 and 1/8 of the frame rate. However, since MCLIFT is an embedded 3D wavelet-based video coder, the encoded bitstream can be scaled in the quality level as well. Such combined scalability is not feasible for MPEG, and is very useful for the Internet applications. The draw back is that MCLIFT doubles the coder delay of MPEG. The memory required to implement the MCLIFT is also 5 frames versus the 3 frames required by the MPEG. Moreover, since the prediction in MCLIFT is based on the original frames, there are local drifting errors from category A frames to category B frames to category C frames to category D frames.

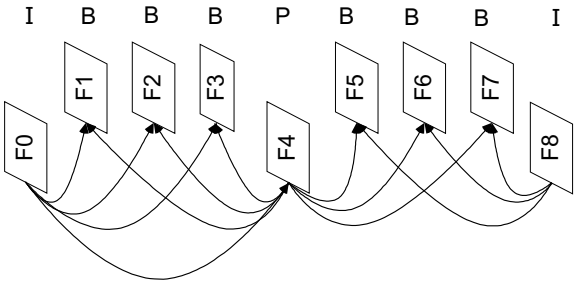

Figure 5 MPEG temporal frame structure.

2.2 Intra-frame wavelet decomposition

Each frame of MCLIFT filtered video is further decomposed by a 3-level bi-orthogonal 9-7 wavelet filter in the horizontal and vertical direction. Only the low-pass of each level is further decomposed. The spatial wavelet decomposition further packs the energy of the video into a small number of wavelet coefficients.

2.3 Block-based Arithmetic Entropy Coding

After MCLIFT and spatial wavelet filtering, the wavelet transformed coefficients are chopped into blocks, where each of the block is encoded independently using an embedded entropy coder. Through embedded coding, the compressed bitstream of each block can be further truncated at a later stage to form a bitstream fully scalable in the frame rate and quality level. The 3D block entropy coder developed in [16] is used in this work. Though the coder in [16] is developed to compress the concentric mosaics, it can be directly used for 3D wavelet video coding.

There are many forms of the embedded entropy coders. In this work, a context-based arithmetic bit-plane coder is adopted. The coder encodes each block bit-plane by bit-plane, from the most significant one to the least significant one. In a certain bitplane, we denote those coefficients that are still coded as '0' as the insignificant coefficient. For the coefficient with at least one non-zero bit in the previous bitplane coding, it is denoted as the significant coefficient. Each bitplane is further scanned three times, resulting three passes, predicted significance, predicted insignificance and refinement. The predicted significant pass encodes the coefficient that is still insignificant in the current bitplane, but has at least one significant coefficient in the 26 neighbors. The predicted insignificant pass encodes the coefficient that is still insignificant and has no significant neighbors. The refinement pass encodes the bit of the significant coefficient. In the predicted significant and predicted insignificant passes, the insignificant bit is encoded with an arithmetic coder using a context derived from the 26 immediate neighbors of the current coefficient. We further group the significance of the 26 neighbors into 10 categories so that the number of contexts is reduced to avoid context dilution. The refinement bit is encoded by the same arithmetic coder with the context calculated on whether the coefficient just becomes significant in the previous bitplane. The technique is a direct extension of the block entropy coder used in JPEG 2000[17]. During the bitplane coding, the coding rate R and distortion D of each pass is recorded at the end of each bitplane. The recorded rate-distortion performance of the block is used in the bitstream assembler.

After all the blocks of coefficients have been encoded, a bitstream assembler is used to optimally allocate the bits among different blocks. The rate-distortion theory indicates that optimal coding performance can be achieved if all blocks operate on the same rate-distortion curve. The functionality of the bitstream assembler is thus to find the common rate-distortion slope of all blocks, and calculate the number of included bits for each block. The MCLIFT bitstream is formed by the truncated block bitstream and the bitstream length of all wavelet coefficient blocks. With such an entropy coding and bitstream assembling strategy, the MCLIFT compressed bitstream can be flexibly scaled at different frame rate and quality level.

3. EXPERIMENTAL RESULTS

We compare the proposed motion compensated lifting (MCLIFT) wavelet video coder with two benchmark coders. The first one is a 3D wavelet coder without motion compensation in the temporal direction. We still use the bi-orthogonal 9-7 filter for intra-frame decomposition. However, in the temporal direction, the video is decomposed by a three level wavelet filter with either 5-3 or 9-7 bi-orthogonal wavelet without motion compensation. The decomposed coefficients are then further entropy encoded in the exact same way as that of MCLIFT. The second benchmark coder is the MPEG-4 VM 16.0. Two MPEG-4 GOP structures are tested. The first one encodes the entire video as a GOP with one I frame followed by all P frames. The second GOP structure consists of 8 frames per GOP, with "I B B B P B B B" frame structure shown in Figure 5. We select this GOP structure because it is the GOP structure most similar to the one used by 3-level MCLIFT wavelet. The Inter_4MV mode and the Overlap mode are disabled in MPEG-4 for fair comparison purpose, as the corresponding modules have not been implemented in MCLIFT due to time constraint. Experiments are performed on the Foreman (in Table 1) and Coastguard (Table 2) sequences. Each sequence is at QCIF resolution (176x144 pixels per frame), with a total of 300 frames at 30 frames per second. We compress the wavelet coefficients of the MCLIFT coder at bitrate of 304.13, 152.06 and 76.03 kbps. The motion vectors in the MCLIFT are differentially predicted and entropy coded with the same VLC coder used in MPEG-4 P frame coding [18]. The combined bitrates of the coefficient and motion coding are listed in row 2 of the table. For comparison, we match the coding rate of the benchmark 3D wavelet coder and MPEG-4 with that of MCLIFT. The PSNRs of the coding result of the video sequence Foreman and Coastguard are listed in Table 1 and Table 2, respectively.

We first compare MCLIFT with the benchmark 3D wavelet coder. Since the entropy coding schemes are exactly the same between the MCLIFT and the 3D wavelet coder, the performance difference is due to the different temporal decomposition schemes. It is shown that MCLIFT outperforms the 3D wavelet

coder by an average of 1.6 and 1.1 dB for the 5-3 and 9-7 filter, respectively. The performance advantage provided by the temporal motion compensated lifting (MCLIFT) is significant.

Table 1 Comparison of the MCLIFT, 3D wavelet and MPEG-4 video coder (Foreman sequence)

Bit rate (kbps) / Algorithms	Foreman (30fps, texture + motion rate) Motion rate is 29.53 kbps		
	333.66 (kbps)	181.59	105.56
MCLIFT	37.34	34.33	31.52
5-3 3D wavelet	35.41	32.52	30.40
9-7 3D wavelet	35.82	32.89	30.74
MPEG-4(IPPP..)	36.24	33.66	31.50
MPEG-4 (IBBBPBBB)	36.40	33.36	29.57

Table 2 Comparison of the MCLIFT, 3D wavelet and MPEG-4 video coder (Coastguard sequence)

Bit rate (kbps) / Algorithms	Coastguard (30fps, texture + motion rate) Motion rate is 19.62 kbps		
	323.75 (kbps)	171.68	95.65
MCLIFT	35.50	32.64	30.32
5-3 3D wavelet	33.54	30.93	29.11
9-7 3D wavelet	34.08	31.73	29.78
MPEG-4(IPPP..)	33.85	31.33	29.33
MPEG-4 (IBBBPBBB)	33.98	31.30	28.50

In the second set of experiments, we compare MCLIFT with MPEG-4 VM 16.0. MCLIFT coder outperforms MPEG-4 with IPPP.. and IBBBPBBB GOP structures by an average of 0.9dB and 1.4dB, respectively. The superior performance of the MCLIFT is due to the better de-correlation in the temporal direction through the MCLIFT filter, the superior embedded entropy coder, and the R-D optimized bitstream assembler. Moreover, the MCLIFT compressed bitstream can be flexibly scaled in the frame rate and quality level, which is a very attractive feature in the video streaming and storage applications.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we implemented a motion compensated lifting (MCLIFT) framework for the 3D wavelet video coder. By using bi-directional motion compensation in each lifting step of the temporal direction, the performance is effectively enhanced. In the future work, we'll improve the framework by using decoded reference frames instead of original one to avoid the local drifting problem addressed in Section 2.1.

5. REFERENCES

[1] B-J. Kim; W.A. Pearlman, "An embedded wavelet video coder using three-dimensional set partitioning in hierarchical trees (SPIHT)", *Data Compression Conference. DCC '97*. Proceedings, pp: 251 –260.

[2] B.-J. Kim; Z. Xiong; W.A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)", IEEE Trans. on *Circuits and Systems for Video Technology*, Vol.: 10 Issue: 8, pp: 1374 –1387, December 2000.

[3] Y.-Q. Zhang, S. Zafar, "Motion-compensated wavelet transform coding for color video compression", *Circuits and Systems for Video Technology*, IEEE Trans. on , Vol: 2 Issue: 3 , Sept. pp: 285 –296.1992

[4] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video", *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 572-689, Sept. 1994.

[5] J. R. Ohm, "Three-dimensional subband coding with motion compensation", *IEEE Trans. on Image Processing*, vol. 3, no. 5, pp. 572-588, Sept. 1994.

[6] J. Y. Tham, S. Ranganath, and A. A. Kassim, "Highly scalable wavelet-based video codec for very low bit-rate environment", *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 1, Jan. 1998.

[7] A. Wang; Z. Xiong; P.A. Chou.; S. Mehrotra, "Three-dimensional wavelet coding of video with global motion compensation", *Proceedings. DCC '99, 1999*. Page(s): 404 –413.

[8] J.Z. Xu, S. Li, Y.-Q. Zhang, "Three-Dimensional Shape-Adaptive Discrete Wavelet Transforms for Efficient Object-Based Video Coding", *IEEE/SPIE Visual Communications and Image Processing* (VCIP 2000), Perth, June 2000.

[9] P. Cheng, J. Li and J. Kuo, "Rate control for embedded wavelet video coder", *IEEE Trans. On Circuit and System for Video Technology*, Vol. 7, No. 4, pp. 696-702, Aug. 1997.

[10] D. Marpe, H. Cycon, "Very low bit-rate video coding using wavelet-based techniques" *Circuits and Systems for Video Technology*, IEEE Trans. on , Vol. 9 Issue 1, pp: 85 -94, Feb. 1999

[11] E. Asbun, P. Salama, E.J. Delp, "A rate-distortion approach to wavelet-based encoding of predictive error frames", *IEEE International Conference on Image Processing (ICIP2000)*, Vol.3,pp. 150-153, Vancouver, Canada, Sept. 10-13, 2000.

[12] D. Blasiak, W.-Y. Chan, "Efficient wavelet coding of motion compensated prediction residuals", Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, Volume: 2 , pp: 287 –291, vol.2, 1998.

[13] M. Wien, "Hierarchical wavelet video coding using warping prediction", *Image Processing, IEEE Int. Conf. on (ICIP2000)*, Vol.3, pp.142-145, Vancouver, Canada, Sept. 10-13, 2000.

[14] W. Sweldens, "The lifting scheme: A new philosophy in biorthogonal wavelet constructions," in *Wavelet Applications in Signal and Image Processing III*, pp. 68-79, Proc. SPIE 2569, 1995.

[15] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps", *J. Fourier Anal. Appl.*, vol. 4, pp. 247-269, 1998.

[16] L. Luo, Y. Wu, J. Li, and Y. -Q. Zhang, "Compression of concentric mosaic scenery with alignment and 3D wavelet transform", *SPIE Image and Video Communications and Processing*, vol. 3974, SPIE 3974-10, San Jose, CA, Jan. 2000.

[17] JPEG 2000 editorial committee, "JPEG 2000 image coding system", ISO/IEC JTC1/SC29/WG1N1646, *Mar*. 2000, Japan.

[18] "MPEG-4 Video Verification Model Version 16.0", ISO/IEC JTC 1/SC29/WG11 N3312, March 2000